Application of Information Technology ■

Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence

Halil Kilicoglu, MS, Dina Demner-Fushman, MD, PhD, Thomas C. Rindflesch, PhD,
Nancy L. Wilczynski, PhD, R. Brian Haynes, MD, PhD

Abstract The growing numbers of topically relevant biomedical publications readily available due to advances in document retrieval methods pose a challenge to clinicians practicing evidence-based medicine. It is increasingly time consuming to acquire and critically appraise the available evidence. This problem could be addressed in part if methods were available to automatically recognize rigorous studies immediately applicable in a specific clinical situation. We approach the problem of recognizing studies containing useable clinical advice from retrieved topically relevant articles as a binary classification problem. The gold standard used in the development of PubMed clinical query filters forms the basis of our approach. We identify scientifically rigorous studies using supervised machine learning techniques (Naïve Bayes, support vector machine (SVM), and boosting) trained on high-level semantic features. We combine these methods using an ensemble learning method (stacking). The performance of learning methods is evaluated using precision, recall and F_1 score, in addition to area under the receiver operating characteristic (ROC) curve (AUC). Using a training set of 10,000 manually annotated MEDLINE citations, and a test set of an additional 2,000 citations, we achieve 73.7% precision and 61.5% recall in identifying rigorous, clinically relevant studies, with stacking over five feature-classifier combinations and 82.5% precision and 84.3% recall in recognizing rigorous studies with treatment focus using stacking over *word + metadata* feature vector. Our results demonstrate that a high quality gold standard and advanced classification methods can help clinicians acquire best evidence from the medical literature.

■ **J Am Med Inform Assoc.** 2009;16:25–31. DOI 10.1197/jamia.M2996.

Introduction

Finding high-quality clinical information in a timely fashion is essential to the successful practice of evidence-based medicine. Evidence-based practice (EBP) is a paradigm that emphasizes the explicit and judicious use of the best research evidence currently available for medical decision-making.[1] The most reliable sources of high-quality evidence for clinicians are systematic reviews of the medical literature that identify, evaluate, and present bottom-line advice extracted from high quality studies on specific medical topics.

To ensure that all critical information is summarized and brought to a clinician's attention, the domain experts who write systematic reviews, practice guidelines, and other secondary resources need to examine all relevant publications, paying particular attention to content and methodological criteria. This human-centered approach generates the best EBP resources for many medical specialties. It is, however, labor intensive and time consuming. Recently, automated knowledge-based approaches and statistical techniques have shown promise in identifying high-quality articles and other types of clinical text classifications.

In this paper, we explore the possibility of automatically recognizing MEDLINE® citations containing scientifically rigorous clinical evidence using supervised machine learning techniques. We are particularly interested in whether and to what extent semantic features extracted from MEDLINE citations using natural language processing improve classification results. We further extend our methods to recognizing MEDLINE citations with rigorous clinical evidence concerning a specific clinical purpose, focusing on treatment or prevention of disease. Our ultimate goal is to support EBP by helping domain experts in evaluating and synthesizing best evidence from the medical literature.

Background

Clinical Query Filters

To find high-quality articles concerning any aspect of medical practice in MEDLINE, Wilczynski and colleagues[2] developed optimal "clinical query filters," each of which is a Boolean combination of indexing terms and metadata and is targeted for high specificity or sensitivity for a particular clinical purpose, such as etiology, diagnosis, prognosis, or treatment. These strategies were adapted for use by the Clinical Queries feature of PubMed. More recently, Wong and colleagues[3] extended these filters to identify qualitative

Affiliations of the authors: Department of Computer Science and Software Engineering (HK), Concordia University, Montréal, QC, Canada; National Library of Medicine (HK, DD-F, TCR), National Institutes of Health, Bethesda, MD; Health Information Research Unit (NLW, RBH), McMaster University, Hamilton, ON, Canada.

Supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Correspondence: Halil Kilicoglu, MS, Concordia University, Department of Computer Science and Software Engineering, 1515 Ste Catherine West, Montréal, QC, H3G 1M8, Canada; e-mail: <h_kilico@cse.concordia.ca>.

Received for review: 09/08/08; accepted for publication: 09/30/08.

studies. These efforts involved critical appraisal and manual annotation of 49,028 MEDLINE records from 161 journals published in 2000. This collection created by highly qualified specialists provides the basis for the study presented in this paper.

Related Research

Identification of articles containing high quality clinical evidence can be viewed as a text classification task. Due to the availability of the OHSUMED collection,[4] classification and categorization of medical text by assigning multiple Medical Subject Headings (MeSH) to a document is relatively well studied. For example, an F_1 score of 0.55 was achieved by using a hierarchical classifier to assign categories with at least 75 training examples in this collection.[5] A support vector machine (SVM) classifier trained on documents represented as a combination of a bag-of-words and a "bag-of-biomedical terms" extracted from document titles using MetaMap[6] achieved an F_1 score (harmonic mean of recall and precision) of 0.60 in assigning 634 disease categories to the documents in this collection.[7]

Wilcox and colleagues[8] used five different machine learning techniques to demonstrate that domain knowledge significantly improves classification of medical text reports. In a comparison of classifiers based on domain knowledge (expert-crafted rules), a combination of domain knowledge and statistical methods (a Bayesian network whose structure was created by experts), and a supervised machine learning method (decision tree) for classification of radiology reports, the domain knowledge system slightly outperformed the other two.[9]

Aphinyanaphongs and colleagues[10] used machine learning to automatically construct filters that identify high-quality, content-specific articles in internal medicine. Drawing on MEDLINE citations abstracted and cited in the American College of Physicians (ACP) Journal Club, they built two corpora, one for treatment and etiology and another for prognosis and diagnosis. Their feature set exploits MeSH indexing terms and publication types assigned by NLM indexers as well as words in the title and abstract of the citation. The classifiers tested by Aphinyanaphongs and colleagues[10] include Naïve Bayes, SVM, and text-specific boosting. The polynomial SVM models performed the best, achieving up to 80% (74–83%) recall and 33% (31–34%) precision, when classifying articles into the treatment specific category. The authors note that training the system on a collection different from the one used to create the clinical query filters for PubMed is a limitation of the study.

SemRep

Medical text lends itself to sophisticated document representation, since significant domain knowledge has been encoded in the Unified Medical Language System® (UMLS®),[11] and automatic text processing applications that exploit this knowledge already exist. One such application is SemRep,[12] a knowledge-based system that extracts semantic predications from MEDLINE titles and abstracts. Semantic predications consist of UMLS Metathesaurus concepts as arguments and UMLS Semantic Network relations as predicates. Processing relies on an underspecified syntactic analysis based on the SPECIALIST Lexicon[13] and MedPost part-of-speech tagger.[14] MetaMap[6] is used to map simple noun phrases to Metathesaurus concepts, and "indicator rules" are used to map syntactic elements to allowable Semantic Network predicates (TREATS, PREVENTS, INHIBITS, etc.). For example, given the sentence in (1), SemRep identifies the semantic predications in (2):

(1) *Usefulness of massive oral nicorandil in a patient with variant angina refractory to conventional treatment.*

(2) *Nicorandil TREATS Patients Angina Pectoris*
Variant PROCESS_OF Patients

We relied on SemRep for automatic extraction of domain knowledge for our document classification experiments.

Our research seeks to answer several questions. First, we test whether the results obtained by Aphinyanahongs and colleagues[10] using a gold standard that contains articles cited in the ACP Journal Club generalize to the case in which training is performed on the collection used to develop the clinical queries. Second, we seek to determine whether combining statistical methods with domain knowledge obtained through deep semantic processing can improve precision without a significant degradation in recall in recognizing scientific rigor. Third, we investigate whether domain knowledge improves the recognition of content-specific rigorous studies, focusing on treatment-related studies. Finally, we discuss how disparate feature sets contribute to identification of high-quality, content-specific clinical evidence.

Methods

Data Collection

We used the test collection that was manually created to develop the clinical query filters for PubMed.[2] This collection consists of 49,028 MEDLINE documents, classified across three dimensions: human health care interest (yes/no), scientific rigor (yes/no) and purpose (etiology, prognosis, diagnosis, treatment or prevention, economic studies, clinical prediction guides and reviews). The scientific rigor label is used only for articles to which a purpose label applies. In this study, we focused on scientific rigor and 'treatment or prevention' content area; of the 49,028 documents, 48,126 are unique and 3,036 (approximately 7%) of these are labeled as being scientifically rigorous. 2,228 of these 3,036 documents (approximately 5%) are labeled as being in the 'treatment or prevention' content area.

We used a subset of the collection for our experiments. Our training set consisted of 10,000 documents (750 rigorous, 9,250 non-rigorous; 561 of 750 rigorous documents labeled as having treatment or prevention focus). The test set included 2,000 documents (200 rigorous and 1,800 non-rigorous; 140 with treatment or prevention focus). The documents were randomly selected from those having at least one MeSH indexing term. The latter requirement explains the higher percentage of rigorous studies in our subset compared to the whole collection.

Machine Learning Methods

As did Aphinyanaphongs and colleagues,[10] we experimented with three supervised machine learning methods: Naïve Bayes, SVM, and boosting. These three classifiers have reportedly worked well with text categorization tasks.[15] We further experimented with an ensemble learning method, stacking,[16] to combine predictions of the above three classifiers.

Given a feature *f*, a Naïve Bayes classifier estimates the probability of class *C* using the training data to estimate *P(f|C)* and predicts the class by applying the maximum a posteriori (MAP) decision rule. Although its assumption of conditional independence between features often does not hold, Naïve Bayes classifier performs remarkably well in practice, including in text classification tasks.[17]

Classifiers of SVM use "kernel" functions to map the input space to a higher dimensional space where a maximal separating hyperplane is constructed. This hyperplane corresponds to a nonlinear boundary in the original input space. Linear and polynomial SVM classifiers have been used successfully in text categorization tasks.

Boosting is based on the intuition that a set of simple classifiers (weak learners) can be combined into a single, highly accurate classifier (strong learner). In an iterative process, boosting weights training instances and places higher weights on more difficult training instances in subsequent iterations. A weak learner is then trained using the weighted data set and added to the final strong learner. By combining weak learners, boosting reduces the learner bias without significantly affecting the variance. One of the boosting-based algorithms, BoosTexter, uses very simple decision rules, called "decision stumps" and has been successfully applied to text categorization.[18]

Stacked generalization (or stacking), introduced by Wolpert,[19] is expected to reduce the classification error rate through a process equivalent to cross-validation in some special cases, or through forming a linear combination of the guesses in other special cases. Our stacking approach combines predictions from lower-level models into a higher-level model using a version of least squares linear regression adapted for classification.[20] This multiple linear regression (MLR) meta-classifier has been shown to outperform other methods of combining classifiers.[20]

We applied these methods to two binary classification tasks: learning models that identify high quality, methodologically rigorous MEDLINE articles, and those that identify rigorous MEDLINE articles focusing on treatment or prevention of disease. We used the Naïve Bayes, polynomial SVM, and boosting with decision stumps classifiers provided in the RapidMiner machine learning and data mining package[21] and our own implementation of the stacking classifier.[22] In training the classifiers on the 10,000 training documents, we split the training corpus into ten sets of equal size and performed 10-fold cross validation to avoid overfitting. We did not attempt parameter optimization and used the default settings for the classifiers, presented in Table 1.

Table 1 ■ Classifier Parameters

Classifier	Parameters used
Naïve Bayes	Smoothing value = 1
Polynomial SVM	C (regularization constant) = 1
	Degree = 2
	Misclassification cost = 0.1
	Maximum # of iterations = 50,000
Boosting	# of boosting iterations = 10

SVM = Polynomial support vector machine.

We performed stacking in two ways. One method involved combining predictions from the three base classifiers applied to a specific feature vector (*feature stacking*). In the second stacking scenario, predictions from different base classifiers applied to disparate, basic feature vectors (see the section below) were combined (*feature-classifier stacking*).

Feature Vectors

To determine the contribution of various feature types to the classification task, we experimented with combinations of five non-overlapping basic feature vectors extracted from documents. The five basic feature types are presented below. We use the words in parentheses to refer to them in the rest of this paper.

1. Words in the title and abstract of a MEDLINE citation (*word*)
2. Metadata from a MEDLINE citation (*metadata*)
3. Semantic predications identified by SemRep (*predication*)
4. The UMLS Metathesaurus concepts extracted from title and abstract of the MEDLINE citation (*entity*)
5. The UMLS Semantic Network relations used in semantic predications (*relation*)

Metadata from MEDLINE are MeSH indexing terms (headings and subheadings) and publication types assigned manually by NLM indexers. A semantic predication extracted from the title or abstract of a MEDLINE citation, such as *Nicorandil TREATS Patients* noted above, has relation *TREATS* and entities *Nicorandil* and *Patients*. While there are 32 possible combinations of these five feature types, we focus on the four combinations given below.

1. *word* + *metadata*[10] (2,000 features)
2. *word* + *metadata* + *entity* + *relation* (3,034 features)
3. *entity* + *metadata* (2,000 features)
4. *entity* + *relation* (1,034 features)

We selected these combinations after several iterations, as they allow assessing the contribution of domain knowledge to the classification task.

The preprocessing steps included creating the basic feature vectors and their combinations as well as feature selection. To create the *word* feature set, we tokenized the titles and abstracts of the training documents, eliminated PubMed stop words,[23] stemmed the remaining words using the Porter stemmer,[24] and removed the stems that occurred fewer than three times in the training corpus. This processing yielded a total of 10,906 features consisting of word stems. We then weighted the stems using information gain measure,[25] to identify the most informative stems and considered only the 1,000 features with the highest weights. Each document was then encoded with the number of occurrences of selected word stems. Although 1,000 was chosen arbitrarily, our experiments confirmed it as near optimal. Doubling the size of the *word* feature vector did not yield any significant improvement in the performance of the learning algorithms.

The highest-ranking 1,000 *metadata* features were obtained in a similar manner. Each MeSH heading/subheading pair was encoded as a single feature. Stemming and stop word elimination were not necessary in this case.

Semantic predications were identified using SemRep and the semantic features (*predication, entity, relation*) were ex-

Table 2 ▪ Evaluation Results on the Held-out Test Set Using Various Combinations of Features in Recognizing Methodologically Rigorous Studies

Feature Vector	Classifier	Precision	Recall	F_1 Score	AUC
word + metadata	Naïve Bayes	13.8%	**97.5%**	0.241	0.819
	Poly. SVM	26.5%	38.5%	0.321	0.783
	Boosting	50.7%	56.0%	0.532	0.828
	Stacking	27.0%	64.5%	0.465	0.804
entity + relation	Naïve Bayes	55.1%	65.0%	0.596	0.792
	Poly. SVM	75.0%	22.5%	0.346	0.906
	Boosting	**81.7%**	44.5%	0.576	0.852
	Stacking	61.2%	67.0%	0.640	0.900
metadata + entity	Naïve Bayes	56.4%	68.0%	0.617	0.850
	Poly. SVM	81.6%	20.0%	0.321	**0.916**
	Boosting	76.3%	58.0%	**0.659**	0.855
	Stacking	65.0%	64.0%	0.645	0.904
word + metadata + entity + relation	Naïve Bayes	17.5%	93.5%	0.295	0.863
	Poly. SVM	36.2%	31.5%	0.337	0.825
	Boosting	76.3%	58.0%	**0.659**	0.855
	Stacking	37.0%	63.5%	0.468	0.855

AUC = area under the curve; SVM = Polynomial support vector machine.

tracted from SemRep output files. As with *metadata*, stemming and stop word elimination were not performed. For the *predication* and *entity* feature vectors, the top 1,000 features were used. On the other hand, no additional feature selection was performed for the *relation* feature vector, as there were only 34 unique relations.

Evaluation

We evaluated and compared the effectiveness of the learning techniques and feature selection on a test corpus of 2,000 documents. We calculated precision, recall, and F_1 score as well as the area under the receiver operating characteristic (ROC) curve (AUC). The ROC data points were plotted and obtained by varying the confidence threshold value between the POSITIVE and NEGATIVE classes. The AUC was calculated as the pessimistic estimation, using only the area of all rectangles under the ROC curve. To measure the statistical significance of performance differences between pairs of different classifiers, we used the pairwise t-test.

Results

We report both the classification results obtained in the ten-fold cross-validation performed on the training set of 10,000 documents and the results obtained on the held-out test set of 2,000 documents that were not used for training.

Recognizing Methodologically Rigorous Studies

In cross validation, the polynomial SVM using the *word + metadata + entity + relation* feature vector performed best in terms of average AUC (0.957) and precision (83.2%). The Naïve Bayes classifier using the *word + metadata* feature vector provided the best average recall (83.9%), while the same classifier using the *metadata + entity* vector gave the best average F_1 score (0.690). Overall, boosting achieved the best balance of precision and recall, and the polynomial SVM yielded the best average AUC. Among feature vectors, the *word + metadata + entity + relation* vector performed best, followed by the *metadata + entity*.

The results obtained with cross validation were not, however, predictive of performance in the held-out test set. Classifier performances were similar to those in cross validation, whereas feature vector performances differed. Focusing on classifiers, the polynomial SVM with the *metadata + entity* vector gave the best AUC (0.916). On the other hand, boosting performed best in terms of precision and F_1 score (81.7% and 0.659, respectively) on this feature vector. As with cross validation, Naïve Bayes using the *word + metadata* feature vector provided the best recall (97.5%), at the expense of low precision (13.8%). In considering features, despite moderately strong results in cross validation, vectors including the *word* feature set performed relatively poorly on the held-out test set, particularly in terms of precision. The *metadata + entity* feature vector achieved the best overall classification results. The *entity + relation* and the *word + metadata + entity + relation* feature vectors are next best and provide comparable results. The former emphasizes precision and overall F_1 score, whereas the latter emphasizes recall. Feature stacking had limited success overall. While it did not yield the best performance in any evaluation metric category, it achieved a better balance between F_1 score and AUC with *entity + relation* and *metadata + entity* feature vectors. For each feature vector scenario, the difference between the performances of pairs of classifiers was found to be statistically significant at the 0.01 level. The evaluation results on the held-out test set are given in Table 2. Figure 1 depicts the ROC curve for the *metadata + entity* feature vector.

Recognizing Rigorous Treatment-related Studies

One of the goals of this study was to further pursue machine learning methods for identifying high-quality, scientifically rigorous treatment-related articles as reported by Aphinyanaphongs and colleagues,[10] using a different, more comprehensive gold standard. To this end, we repeated the experiments described by Aphinyanaphongs and colleagues[10] using the *word + metadata* feature vector and our

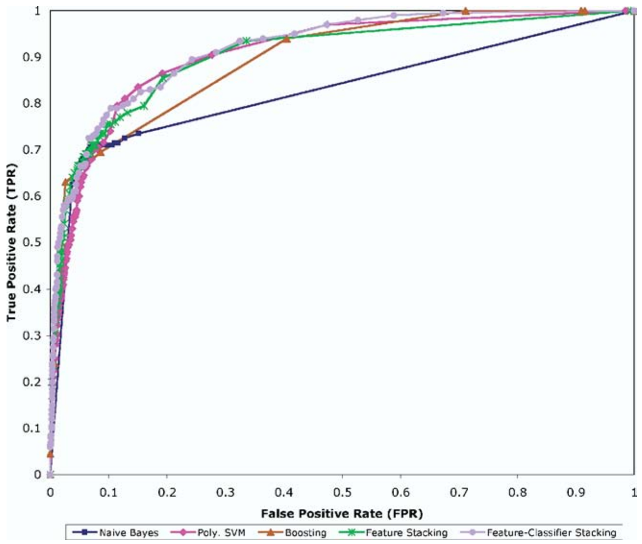

Figure 1. ROC curve for scientifically rigorous studies: base classifiers and feature stacking use the best feature vector (*metadata + entity*) and the feature-classifier stacking uses the combination *metadata(NB) + entity(SVM) + prediction(NB) + relation(SVM) + relation(B)*.

data collection. The results differed significantly, suggesting limited portability of the learning models. Naïve Bayes (0.978), rather than polynomial SVM (0.962), performed best in terms of average AUC in 10-fold cross validation. On the other hand, boosting resulted in an AUC of 0.968, while it provided the best results in terms of F_1 score (0.695), followed by Naïve Bayes (0.641) and polynomial SVM (0.430). The results obtained on the held-out test set were relatively consistent with those obtained in cross validation. Boosting, rather than Naïve Bayes, provided the best AUC and precision in addition to F_1 score, while the Naïve Bayes classifier yielded the best recall. The performance difference between pairs of classifiers was found to be significant at the 0.01 level.

In addition to the *word + metadata* feature vector, we experimented with the three feature vectors that were also exploited for identifying rigorous articles. In cross validation, in terms of AUC and recall, Naïve Bayes with the *word + metadata* feature vector was superior to other feature combinations and base classifiers. On the other hand, the *word + metadata + entity + relation* feature combination with polynomial SVM provided the best precision (80.1%) and achieved the highest F_1 score of 0.727 with boosting. The results obtained in cross validation were predictive of those obtained on the held-out test set; best recall and AUC were again obtained with the *word + metadata* feature vector, latter with the boosting classifier. The *word + metadata + entity + relation* feature vector with boosting yielded the best F_1 score (0.773), while it was outperformed by the combination of polynomial SVM classifier and the *metadata + entity* vector in terms of precision (86.5%).

Feature stacking was more successful in finding rigorous treatment-related studies. In all four scenarios, AUC improved compared to the best performing classifier, whereas F_1 score improved in all but one scenario. On the other hand, precision and recall are almost always lower than those achieved by the best performing classifier. The differences between the performances of classifier pairs for each feature vector were found to be statistically significant at the 0.01 level. The results on the held-out test set in identifying treatment-related articles are presented in Table 3. The ROC curve regarding the best feature vector (*word + metadata + entity + relation*) is given in Figure 2.

Table 3 ■ Test Set Evaluation of Classifiers Trained on Various Feature Vectors in Recognizing Rigorous, Treatment-related Articles

Feature Vector	Classifier	Precision	Recall	F_1Score	AUC
word + metadata	Naïve Bayes	53.2%	88.6%	0.665	0.967
	Poly. SVM	73.5%	25.7%	0.381	0.962
	Boosting	76.9%	71.4%	0.741	0.976
	Stacking	82.5%	84.3%	0.834	0.978
entity + relation	Naïve Bayes	55.7%	80.7%	0.659	0.927
	Poly. SVM	81.6%	28.6%	0.423	0.947
	Boosting	79.4%	57.9%	0.669	0.941
	Stacking	72.5%	71.4%	0.719	0.970
metadata + entity	Naïve Bayes	54.5%	86.4%	0.669	0.965
	Poly. SVM	**86.5%**	22.9%	0.361	0.961
	Boosting	79.1%	72.9%	0.758	0.960
	Stacking	68.0%	83.6%	0.750	0.970
word + metadata + entity + relation	Naïve Bayes	52.5%	**91.4%**	0.667	0.961
	Poly. SVM	81.3%	27.9%	0.415	0.965
	Boosting	82.3%	72.9%	0.773	0.972
	Stacking	74.7%	86.4%	**0.801**	**0.983**

AUC = area under the curve; SVM = Polynomial support vector machine.

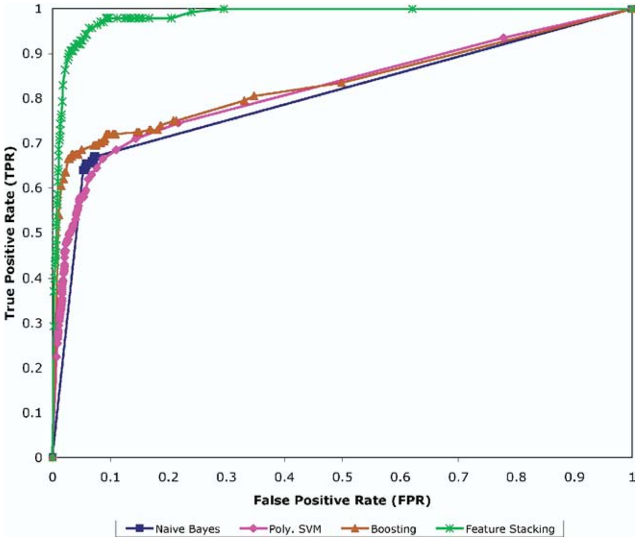

Figure 2. ROC curve in identifying treatment-related scientifically rigorous studies: the base classifiers and the stacking classifier use the best feature vector (*word + metadata + entity + relation*).

Stacking with Different Feature Vectors (Feature-classifier Stacking)

Stacking can be applied to predictions from different classifiers trained on a single feature vector. Another possibility is to use stacking to combine predictions from different classifiers trained using different features. Low performance of the first type of stacking in recognizing scientifically rigorous studies motivated us to further explore the second option. For this purpose, we trained basic feature vectors separately on the three base classifiers and exhaustively stacked predictions from all combinations of classifiers and feature vectors. Evaluation results regarding the combinations that achieve the best performance in one of the evaluation metrics on the held-out test set are presented in Table 4. The last row of Table 4 shows the feature vector-classifier combination that achieves the best balance of F_1 score and AUC. The combination that provided the best AUC (first row) is also depicted in the ROC curve in Figure 1.

In general, feature-classifier stacking improves over feature stacking. All but one (fourth row) of the combinations shown in Table 4 achieve better AUC-F_1 score balance than that achieved by the best base classifier (*metadata + entity* feature vector with boosting).

Discussion

Our preliminary results confirm findings of Aphinyanaphongs and colleagues[10] that machine learning approaches can be used to recognize the methodologically rigorous articles that form the basis of evidence-based medicine, using a well-constructed, comprehensive collection of annotated documents and sophisticated document representa-

Table 4 ■ Evaluation Results for Stacking with Various Feature-classifier Combinations

Feature Vector − Classifier Combination	Precision	Recall	F_1 Score	AUC
metadata(NB) + entity(SVM) + predication(NB) + relation(SVM) + relation(B)	73.0%	58.0%	0.646	**0.919**
word(SVM) + metadata(NB) + entity(NB) + entity(SVM) + entity(B)	73.7%	61.5%	**0.670**	0.892
metadata(NB) + entity(NB) + entity(SVM) + relation(SVM) + relation(B)	**79.6%**	56.5%	0.661	0.912
word(NB) + metadata(B) + entity(B) + predication(SVM) + predication(B)	26.0%	**88.0%**	0.401	0.819
metadata(NB) + entity(NB) + entity(SVM) + predication(SVM) + relation(B)	72.1%	62.0%	0.667	0.908

NB = Naïve Bayes; SVM = Polynomial support vector machine; B = boosting.

tion. Furthermore, these results indicate that machine learning methods can also help to identify scientifically rigorous studies with a more specific treatment or prevention focus.

Scientific Rigor

In recognizing scientifically rigorous articles in general, the *word* features were less effective than the *metadata* and *entity* features. The *word* features performed fairly well in cross validation; however, they were less successful than expected on the test set, indicating that the training and test sets may have different word characteristics or that the size of the test set may be too small. The entity features essentially address synonymy and multi-word expressions in the medical domain by normalizing terms to UMLS Metathesaurus concepts, and our results demonstrate that normalization benefits the recognition of rigorous articles. Similarly, the *metadat*a features, particularly manually assigned MeSH indexing terms, are a higher-level representation of article content and perform comparably to the *entity* features, benefiting the classification task. The effectiveness of the *metadata* and *entity* features most probably stems from their ability to capture and standardize the essence of the scientifically rigorous studies. The *relation* features are few and they contribute slightly but positively to overall performance. On the other hand, the *predication* features, while yielding the best precision among feature vectors consisting of a single feature type when used with boosting (not shown), often have a detrimental or very little effect on performance, perhaps due to the underspecified approach of SemRep to text analysis. However, to a limited extent, it seems to contribute to stacking, as shown in Table 4. Not surprisingly, *metadata* + *entity* feature vector combination achieves the best overall results.

Regarding the base classifiers used in identifying methodologically rigorous studies, boosting consistently strikes the best balance between precision and recall, whereas Naïve Bayes in general performs well on recall (demonstrating a tradeoff between recall and precision), as does polynomial SVM on precision. The AUC results are mixed, although boosting has a slight edge overall. These results demonstrate that different classifiers can be used to satisfy different information needs (SVM for specificity, Naïve Bayes for sensitivity, and boosting for balance between the two, for example).

Feature stacking seems to have an averaging effect over the predictions of the base classifiers, making it potentially useful in terms of achieving a sensitivity-specificity balance, similar to boosting. On the other hand, feature-classifier stacking improves over base classifier performance, demonstrating the value of combining models learned on disparate feature vectors over naively combining disparate features into a single large feature vector for supervised learning.

Treatment-related Scientific Rigor

Based on the commonly used *word* + *metadata* feature set on a different gold standard, we obtained results significantly different than those reported by Aphinyanaphongs and colleagues[10] for recognizing treatment studies. Boosting and, to an extent Naïve Bayes, outperformed the polynomial SVM classifier, reportedly yielding the best performance. Considering that ACP Journal Club inclusion criteria are the same as those used in creating our gold standard, these results may be due to a larger ratio of POSITIVE examples in our training set (5.6% vs. 2.4% in Aphinyanaphongs et al.[10]).

Regarding features in recognizing high-quality treatment studies, *word* features had more success than in recognizing rigorous studies in general. On the other hand, *entity* and m*etadata* features again contribute the most to the classification task, whereas *relation* features provide slight improvement and *predication* features have little positive effect. Overall, *word* + *metadata* + *entity* + *relation* achieves the best performance, although *metadata* + *entity* provides comparable performance with fewer features, confirming the dominance of *metadata* and *entity* features.

Among machine learning techniques, boosting achieves the best performance in terms of F_1 score, while Naïve Bayes consistently gives the highest recall in each scenario. Results with polynomial SVM are mixed; it yields the lowest recall consistently, while achieving precision similar to that achieved with boosting. As in scientific rigor recognition, there is no clear winner in terms of AUC, with boosting being slightly better than the other two classifiers. These results confirm the conclusion that we reached regarding rigorous study recognition: that each base classifier satisfies a different information need. In identifying treatment-related rigorous studies, stacking improves classification performance significantly, even with feature stacking, which provided little improvement in recognizing rigorous studies in general.

Significantly better results obtained in recognizing high quality treatment-oriented studies confirm the findings of Aphinyanaphongs and colleagues,[10] who had less success in identifying diagnosis-, prognosis- and etiology-related studies. This is partly due to relatively small number of studies in the gold standard related to these clinical aspects.

Overall, the results demonstrate that sophisticated document representation using domain knowledge encoded with UMLS Metathesaurus concepts and UMLS Semantic Network relations, automatically extracted using natural language processing, as well as MeSH indexing terms and publication type manually added to the documents provide considerable value for classification effectiveness.

Our supervised machine learning classification approach depends on manual annotation performed by domain experts. Although we demonstrate successful reuse of the collection created independently and prior to our study, it would be highly desirable to develop alternative methods, for example, weakly supervised machine learning, active learning, and dynamic selection of documents for annotation. These methods have a potential to achieve equivalent classification results using small training sets which would significantly reduce the annotation effort.

Future Work

We are interested in extending recognition of high-quality, treatment-related studies to other clinical purposes, such as diagnosis and prognosis. Exploiting the entire gold standard will be instrumental in this research.

Our empirically determined feature reduction was near optimal; however, more theoretically sound dimensionality reduction techniques will be employed in the future. We did not focus on optimizing the classifier parameters in this study and simply used parameters that led to high accuracy in previous studies. For instance, parameter optimization combined with the choice of a good kernel may further improve performance of an SVM classifier.

Conclusion

Using a hand-annotated database of MEDLINE citations, we conducted experiments exploring the role of higher-level semantic features in supervised machine learning techniques for identifying rigorous scientific studies to support evidence-based practice. To the best of our knowledge, this is the first exploration of higher-level semantic features in identification of scientifically rigorous studies. The high level semantic features, particularly entities identified with natural language processing, had a significant positive effect on classification results. Manually assigned metadata improved classification effectiveness, as well. In addition, we show that combining commonly used classifiers and disparate features in various ways using stacking further improves recognition of rigorous studies. We demonstrate that the high quality set of annotated documents and advanced supervised classification methods support a system that shows considerable promise in helping clinicians acquire best evidence from the medical literature.

References ■

1. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. Edinburgh: Churchill Livingstone; 1998.
2. Wilczynski NL, Morgan D, Haynes RB and the Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decision Making 2005;5:20.
3. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. Medinfo 2004;11:311–6.
4. Hersh W, Buckley C, Leone T, Hickman D. OHSUMED: an interactive retrieval evaluation and new large text collection for research. Proc SIGIR 94. 1994;192–201.
5. Ruiz ME, Srinivasan P. Hierarchical neural networks for text categorization. Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 281–282, 1999.
6. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 2001;17–21.
7. Yetisgen-Yildiz M, Pratt W. The effect of feature representation on MEDLINE document classification. Proc AMIA Symp 2005; 849–53.
8. Wilcox A, Hripcsak G, Friedman C. Using Knowledge Sources to Improve Classification of Medical Text Reports. Proc Workshop on Text Mining (KDD-2000). Boston MA 2000.
9. Chapman WW, Fiszman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. J Biomed Inform 2001; 34:4–14.
10. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text Categorization Models for High-Quality Article Retrieval in Internal Medicine. J Am Med Inform Assoc. 2005;2:207–16.
11. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281–91.
12. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36(6):462–77.
13. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annual Symposium on Computer Applications in Medical Care 1994; 235–9.
14. Smith L, Rindflesch TC, Wilbur WJ. MedPost: a part-of-speech tagger for biomedical text. Bioinform 2004;20(14):2320–1.
15. Sebastiani, F. Machine Learning in Automated Text Categorisation. ACM Computing Surveys 2002;34(1):1–47.
16. Ting KM, Witten IH. Issues in stacked generalization. J Artif Intell Res 1999;10:271–289.
17. McCallum A, Nigam K. A Comparison of Event Models for Naïve Bayes Text Classification. In AAAI/ICML-98 Workshop on Learning for Text Categorization. 1998;41-48.
18. Schapire RE, Singer Y. Boostexter: A Boosting-based System for Text Categorization. Machine Learning. 2000;39(2/3):135–168.
19. Wolpert DH. Stacked generalization. Neural Networks. 1992; 5(22):241–59.
20. Ting KM, Witten IH. Issues in stacked generalization. J Artif Intell Res. 1999;10:271–89.
21. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. YALE: Rapid prototyping for complex data mining tasks. Proc SIGKDD 2006;935–40.
22. Demner-Fushman D, Few B, Hauser SE, Thoma G. Automatically identifying health outcome information in MEDLINE records. J Am Med Inform Assoc. 2006 Jan-Feb;13(1):52–60.
23. Princeton University Fine Hall Biology Library, Medline Database: Stop Words, 10/31/08. Available at: http://biolib.princeton.edu/instruct/MedSW.html. Accessed October 2008.
24. Porter MF. An algorithm for suffix stripping. Program 14(3): 130–7.
25. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. Proc 14th International Conference on Machine Learning. 1997;412–20.